DYNAMICALLY CATEGORIZING ENTITY INFORMATION

Cross-Reference to Related Applications

This application claims the benefit of United States Provisional Application Serial No. 60/097029 entitled "Collecting, Combining, Analyzing, and Using Internet and Business Information" filed on August 17, 1998, which is incorporated herein.

Background of the Invention

This application relates to dynamically categorizing entity information.

On the World-Wide Web ("Web"), services known as Internet or Web portals provide hierarchical directories of Web sites. The hierarchical directories include, in Web pages organized by categories, links to Web sites and home pages that are under the control of entities such as businesses. Typically, both the creation of the categories and the assignment of Web sites to the categories are performed with substantial human input, as are any changes to the categories and assignments. For example, after a medical category is created for medical entities such as hospitals, a human searches for hospitals that have Web sites and then assigns the Web sites to the medical category. In such a case, if the medical category is then broken up into multiple subcategories such as a small hospitals subcategory and a large hospitals subcategory, a human must reassign the hospital Web sites to the proper subcategories, by determining which hospitals

qualify as small hospitals and large hospitals and reassigning each of the corresponding hospital Web sites accordingly.

Much of the information available on Web sites is organized into Web pages that can be retrieved and displayed by Web browser software under the direction of a user. Each of the Web pages is identifiable by a respective Uniform Resource Locator text string ("URL"), such as "http://www.isp321.com/frontpage.html", that the Web browser software can use to select the page. Each URL includes a domain name, such as "isp321.com", that identifies the Web site where the corresponding Web page is stored for retrieval by Web browser software. Each domain name is registered by an entity that controls the corresponding Web site and Web pages. A domain name registry organization maintains the domain name registration information, which may include name, address, and other information that allows the organization to bill the entity for payment for the maintenance. (It is to be understood that the term "registry", as used herein, also refers to a domain name registrar or any other entity that may provide assistance in registering a domain name.)

An Internet service provider ("ISP") is an example of an entity that may have a registered domain name for a Web site. Typically, an ISP has customers such as individuals or businesses for whom the ISP stores Web pages on the Web site for retrieval by Web browser software. For example, the ISP may have a

customer Maple Street Plumbing for which the ISP stores a home Web page having a URL that includes a prefix "http://www.isp321.com/~maplestplumb". A home Web page is typically the only or the primary entry point into a Web site or a set of Web pages that are under the control of an entity.

A Web portal is another example of an entity that may have a registered domain name. Typically, a Web portal site allows another entity to create a link from the Web portal site to the other entity's Web site or home page by submitting information to the Web portal site.

Some information about an entity may not be available on a Web site that is under the control of the entity. For example, public financial information about a company may be stored in a database that is not linked to the company's Web site or is not directly accessible by Web browser software, such as a database under the control of a financial services firm.

Summary of the Invention

A method and a system are provided that allow categorized directories of Web sites to be created, maintained, and reconfigured easily without excessive human intervention, and that allow the Web sites listed in the categorized directories to be associated with links to additional information about the respective entities that have control over the Web sites. A set of criteria (such as

geographical location or corresponding standard industry code) is acquired, from a user or elsewhere, that defines a category of entities. The set of criteria is dynamically applied, to a source such as an entity information database, to identify an entity that meets the criteria. It is determined, from a domain name registration organization or an ISP or elsewhere, that the entity is registered as having control over at least a portion of a World-Wide Web address. The at least a portion of a World-Wide Web address is associated with the entity in a presentation, such as a Web page, that indicates that the entity meets the set of criteria (that is, it belongs to the category of entities). In the presentation, a link may be included to a set of computer data about the entity, such as information about the entity in the entity information database. The set of computer data includes information other than information provided at the World-Wide Web address.

Other features and advantages will become apparent from the following description, including the drawings, and from the claims.

Brief Description of the Drawings

Figs. 1-4 are block diagrams of computer-based systems.

Fig. 5 is a flow diagram of a computer-based procedure.

Figs. 6-7 are illustrations of output produced by software.

Figs. 8-9 are illustrations of database information.

Fig. 10 is an illustration of computer file information.

Detailed Description

Fig. 1 illustrates a computer system 10 in which a mapping database 12 maps URLs or domain names 14 to entities 16 such as people, businesses, or government agencies, as described in more detail below. For example, the mapping database may indicate that any URL that begins with "http://www.uspto.gov" is for a Web page controlled by the U.S. Patent and Trademark Office, or that domain names "elmstdogs.com" and "elmstcats.com" are under the control of a company named Elm Street Pets, Inc.

Numerous applications, such as the directory application described below, can take advantage of the mapping database.

Fig. 2 illustrates a computer system 20 having the mapping database, a search engine 22, a Web page record database 24 that includes Web page records 26a-26d, and an entity information database 28 (also known as a business data database) that includes information such as geographic information about entities to which URLs or domain names are mapped in the mapping database.

The mapping database may use a unique identification number ("unique ID"), such as a 9-digit American Business Information ("ABI") number, to identify

an entity so that other information about the entity can be retrieved from the entity information database or elsewhere by searching under the unique ID. (ABI numbers are sponsored by infoUSA.) For example, unique IDs from the mapping database may be used to search the entity information database to produce a subset of the mapping database that has records only for entities having a particular characteristic, such as hospitals having a particular geographic location or more than 1000 employees.

Where an entity constitutes a portion of another entity, each of the entities may be assigned different unique IDs, and the different unique IDs may be linked in the mapping database to note the relationship among the entities. For example, a company that has offices in different locations may be assigned a unique ID for the company itself and a respective different unique ID for each location. In another example, when two previously unrelated companies merge or one is acquired by the other, each may retain its unique ID and a new, different unique ID may be assigned to the combination of the two companies, or both companies may be assigned the same unique ID.

In the entity information database, each entity may be associated with standard industry code ("SIC") fields for SIC numbers that indicate the industry categories for the entity. Each SIC field may be arranged to hold a number having sections to indicate broad and narrow industry categories. For example,

the SIC field may hold an SIC number having six digits, of which the first two digits may indicate a broad industry category such as "service companies", the second two digits may indicate "computer service companies" as a subcategory of "service companies", and the third two digits may indicate "manufacturer computer service companies" as a subcategory of "computer service companies". As a result, the entity information database can be searched by industry categories or subcategories represented by SIC numbers.

Information in the mapping database may be derived from information submitted by or on behalf of the entity when a domain name is registered. For example, when the company Elm Street Pets, Inc. registers the domain names "elmstdogs.com" and "elmstcats.com" with a domain name registry, the company associates the domain names with at least enough information, such as name, address, and telephone number information, to allow the domain name registry to bill the company for maintenance of the registration.

The entity may submit information to the mapping database in other ways such as in an on-line questionnaire that feeds the mapping database.

Information in the mapping database may be derived from information provided by an intermediary such as an ISP or an Internet portal. For example, an ISP having a domain name "isp321.com" may have a customer Maple Street Plumbing for which the ISP hosts and administers a home page having a home

page address "www.isp321.com/~maplestplumb". In such a case, the ISP may have name, address, and telephone number information for the purpose of billing Maple Street Plumbing for such hosting and administration, and may allow such information along with the home page address to be used to link the home page address to Maple Street Plumbing in the mapping database.

In another example, an Internet portal may allow an entity such as Maple Street Plumbing to create an entry or listing named "Maple Street Plumbing" in a "plumbing" section of a on-line directory maintained by the portal, to allow a user to view home page "www.isp321.com/~maplestplumb" by selecting the entry. In such a case, the Internet portal may allow information in the entry, and perhaps any address and telephone number information submitted by the entity during creation of the entry, to be used to link the home page to Maple Street Plumbing in the mapping database.

The mapping database and applications based on the mapping database may take advantage of a hierarchical organization of Web pages, by treating similarly a mapped page and all pages below the mapped page, such as pages sharing a particular prefix with the mapped page. For example, all pages sharing the prefix "www.isp321.com" may be treated as being under the control of an ISP named Global ISP Co. Since such pages include pages sharing the prefix "www.isp321.com/~maplestplumb", which should be treated as being under the

control of Maple Street Plumbing, execution of a unique ID tagging procedure for Global ISP Co. should be followed by execution of a unique ID tagging procedure for Maple Street Plumbing so that tags referring to Global ISP Co. are changed to tags referring to Maple Street Plumbing where appropriate.

The mapping database may map an entity to Web pages maintained at different Web sites. For example, Maple Street Plumbing may have a first set of Web pages at the Global ISP Co. site and a second set of Web pages at another ISP's site.

The entity information database may include a database such as EDGAR that includes information about companies. Information derived from EDGAR may be used to allow a search of the entity information database to be limited to companies that match a specified financial profile, such as profitable companies.

Information in the mapping database or the entity information database may allow searches to be limited by relative size of entities, such as size in an industry.

One or more of the databases referenced above may be or include a relational database and may have records to which fields may be added readily to accept informational tags and Web link information.

Fig. 3 illustrates an example of a directory application system 400 in which information drawn from the mapping database 12 and the entity information

database 28 is used to produce a categorized directory view 402 of entities and Web sites that are under the control of the entities. The entity information database has information about attributes of each entity, such as the entity's location, a SIC code for the entity, and the size of the entity by the number of employees. In the categorized directory view, the entities and the Web sites under the control of the entities are grouped according to one or more of the attributes. For example, the grouping may be according to subject matter areas identified by SIC codes, or according to business status such as privately-held, public, or not-for-profit.

In at least the case of an example embodiment 500 described below, a directory application 404 draws information from the mapping database and the entity information database at the time a categorized directory view is produced, so that the information presented in the view is as current as the information in the mapping and entity information databases. The example embodiment described below also provides a highly data storage space efficient implementation that does not require an intermediate database of directory contents; the directory application relies on a category file 406 that defines the categorical structure, but not the contents, of the view.

With reference to Figs. 4-5, example embodiment 500 is now described in connection with a procedure 1000. A category file 502 is maintained that defines a

hierarchical categorical structure of SIC codes (step 1010). As shown in an example in Fig. 10, the category file has at least a top level of broad subject matter categories, and may also have lower levels of subcategories. Each category and subcategory is associated with a respective set of SIC codes.

A top level directory view is presented (step 1020). The top level directory view lists each of the categories in the top level, and may also allow the user to specify geographical criteria for a geographical filter, which filter is used as described below. Fig. 6 illustrates an example of a top level directory view.

The user is allowed to select a lowest level category or subcategory (step 1030). If there are no subcategories, a top level category may serve as the lowest level category. Otherwise, the user can burrow down through one or more levels of subcategories until one of the lowest level subcategories is reached. The lowest level category is selected by one or more mouse clicks or other user entry, and hierarchical views showing categories and subcategories are presented as necessary to help the user navigate.

It is determined, from the category file, which SIC codes are associated with the lowest level category that was selected ("selected SIC codes") (step 1040).

The entity information database is searched to acquire information indicating which entities match the selected SIC codes ("selected entities") (step

1050). The selected entities information may identify the selected entities by their respective unique IDs.

A view 504 is presented in which the selected entities are listed under the selected lowest level category (step 1060). As the selected entities are listed, the mapping database is searched to determine URLs for Web sites that are under the control of the selected entities ("selected URLs") (step 1070), and the selected URLS are listed together with the respective selected entities (step 1080). Fig. 7 illustrates an example of view of the list of the selected entities. From the view, the user can retrieve a page from a selected entity's Web site by selecting the respective selected URL (e.g., "www.nationalparks.org"), and can retrieve information about the entity from the entity information database by selecting the name of the entity in the list (e.g., "US Interior Dept").

In alternative embodiments, before or as the view of the list of the selected entities is presented, the selected entities are filtered, e.g., to filter out selected entities that do not match geographical criteria specified by the user (step 1090), or that are indicated in the mapping database as having control over no Web sites (step 1100). The geographical filter may be implemented by including the geographical criteria with the selected SIC codes when the entity information database is searched to determine selected entities. For example, if the user's geographical criteria specify a city, the only entities that are included in the view

of the list of the selected entities are entities that are indicated in the entity information database as having a primary or secondary address in the city. In the user's geographical criteria, the presence may be required to be a primary address, and the directory application may allow this requirement to be selected by an input selection prompt such as box 700 of Fig. 7.

The directory application or another application such as a search engine may keep track of the number of times a Web site is accessed (i.e., the site's popularity), and may sort the list of selected entities by the number of times the Web sites under the control of the respective entities have been accessed, or the category file may be structured to group entities by the number of times their web sites have been accessed.

Multiple directory applications based on different category files may draw information from the same mapping database and the same entity information database, and thereby may be updated effectively instantaneously and simultaneously as the mapping and entity information databases are updated. For example, a first directory application may be provided to group entities by SIC codes as described above, and a second directory application may be provided to group entities by relationship to medical specialties or other professional specialties. In such a case, if the mapping database is updated to reflect that a medical office has made a change to the Web address of the Web site that is

under the control of the medical office, the change is included in any views that are subsequently presented by the first and second directory applications and that list the medical office as a selected entity.

Other information that may be used in a category file to group entities includes brands of goods or services that may be offered by or through entities, and information listed in Figs. 8-9 that illustrate an example of information that may be stored in the entity information database for each entity.

The user may also be permitted, in accordance with an input mechanism such as box 702 of Fig. 7, to execute a Web site search limited to Web sites identified by the selected URLs. Such a search may be accomplished by directing search engine 22 (Fig. 2) to search Web page record database 24 and then filtering out of the search results any records that do not include at least one of the selected URLs. Alternatively, if the records in the Web page record database have been tagged with unique IDs corresponding to entities, such a search may be accomplished by retrieving, from the entity information database, the unique IDs for the selected entities ("selected unique IDs"), and directing the search engine to return only records that are tagged with one or more of the selected unique IDs.

Any of many different types of computer equipment may be used. For example, one or more Intel-based personal computers may be used that run an

SQL database on Linux and one or more programs written in Perl or the C programming language with interfaces to the SQL database.

The technique (i.e., the procedures described above) may be implemented in hardware or software, or a combination of both. In at least some cases, it is advantageous if the technique is implemented in computer programs executing on one or more programmable computers, such as a personal computer running or able to run an operating system such as Unix, Linux, Microsoft Windows 95, 98, or NT, or MacIntosh OS, that each include a processor, a storage medium readable by the processor (including volatile and non-volatile memory and/or storage elements), at least one input device such as a keyboard, and at least one output device. Program code is applied to data entered using the input device to perform the technique described above and to generate output information. The output information is applied to one or more output devices such as a display screen of the computer.

In at least some cases, it is advantageous if each program is implemented in a high level procedural or object-oriented programming language such as Perl, C, C++, or Java to communicate with a computer system. However, the programs can be implemented in assembly or machine language, if desired. In any case, the language may be a compiled or interpreted language.

In at least some cases, it is advantageous if each such computer program is stored on a storage medium or device, such as ROM or optical or magnetic disc, that is readable by a general or special purpose programmable computer for configuring and operating the computer when the storage medium or device is read by the computer to perform the procedures described in this document. The system may also be considered to be implemented as a computer-readable storage medium, configured with a computer program, where the storage medium so configured causes a computer to operate in a specific and predefined manner.

Other embodiments are within the scope of the following claims. For example, the user may be a human being or a non-human entity such as a computer program or an automated device that may interact with one or more of the databases or one or more of the applications via an application programming interface ("API") or a network message. An on-line information store or multiple databases may serve as the entity information database, which may take the form of any mechanism that provides automated access to information, such as a spreadsheet file or a store of email messages.